(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) EP 1 343 351 A1

# (12) EUROPEAN PATENT APPLICATION

(43) Date of publication:
**10.09.2003 Bulletin 2003/37**

(51) Int Cl.[7]: **H04R 3/00**, H04R 1/40, G10K 11/178

(21) Application number: **02388021.4**

(22) Date of filing: **08.03.2002**

(84) Designated Contracting States:
**AT BE CH CY DE DK ES FI FR GB GR IE IT LI LU MC NL PT SE TR**
Designated Extension States:
**AL LT LV MK RO SI**

(71) Applicant: **TELEFONAKTIEBOLAGET LM ERICSSON (publ)**
**126 25 Stockholm (SE)**

(72) Inventors:
- **Grbic, Nedelko**
  **37 132 Karlskrona (SE)**
- **Nordholm, Sven**
  **WA 6148 Shelley (AU)**
- **Claesson, Ingvar**
  **240 10 Dalby (SE)**
- **Lindgren, Ulf**
  **226 49 Lund (SE)**

(74) Representative: **Sigh, Erik et al**
**Zacco Denmark A/S**
**Hans Bekkevolds Allé 7**
**2900 Hellerup (DK)**

## (54) A method and an apparatus for enhancing received desired sound signals from a desired sound source and of suppressing undesired sound signals from undesired sound sources

(57) In a method of enhancing received desired sound signals such as speech signals, and of suppressing undesired sound signals such as noise. The method uses I microphones each feeding into a bank of K band pass filters. In the I identical banks of K band pass filters each bank receives an input from one microphone and has K outputs of distinct frequency sub-bands. K beam-formers, one for each frequency sub-band, each receives one input from each of the I filter banks representing the same frequency sub-band. The output from each beam-former is the beam-formed signal for one distinct frequency sub-band. From the beam-formed frequency sub-band signals a time domain output signal is reconstructed. Stored estimates of the desired sound source are used in the method of the invention. Such estimates are obtained from received signals from the desired sound source at times with no or insignificant undesired sound signals. The method lends itself in particular to hands-free mobile communication in noisy environments such as in a motor vehicle.

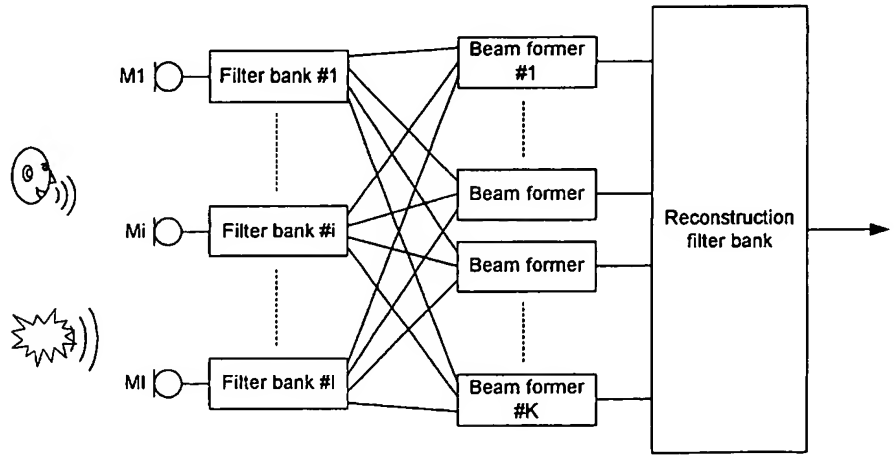


Fig. 1

EP 1 343 351 A1

## Description

### Technical field of the invention

**[0001]** This invention relates to enhancing of desired sound signals such as speech signals and to suppression of noise and interfering sound sources by using an array of microphones and adaptive beam forming.

### Background of the invention

**[0002]** In traditional verbal communication such as telephone communication a single microphone is located near the speaker's mouth. The microphone can be part of a handset or of a headset. In hands free verbal communication the microphone or microphones will usually be in a position more remote to the speaker's mouth than in handheld communication. Hands free communication includes the situations where eg a telephone handset or a microphone is placed on a conference table for picking up the speech of several speakers, and where a microphone is mounted in a car for picking up the speech of a person in the car, typically of the driver. In hands free communication the microphone will usually be placed in a position more remote from the speaker's mouth than in handheld communications, whereby the received speech signal from the speaker will be weaker, which in turn reduces the signal-to-noise ratio, and the intelligibility of the speech will be reduced.

**[0003]** With pure frequency or narrow frequency-band signals from a fixed point source beam forming can be performed with a set of microphones so arranged in space that the sum of the microphone signals will have a maximum response at the source. Furthermore, by individually delaying the microphone signals the point of maximum response can be moved electronically.

**[0004]** If the point source emits a wide frequency-band signal such as a speech signal, the beam former must be able to delay a plurality of frequencies individually. This is called wide frequency-band beam forming, and signals from a certain location are allowed to pass the system, whereas signals outside that location are attenuated or even cancelled. A common way of achieving this is to use digital linear filters at each microphone signal.

**[0005]** In general, wide frequency-band beam forming methods make use of fundamental properties of the spatial and/or the temporal distribution of both the speech source and the noise sources in order to improve the speech signal quality. Roughly speaking, beam-forming methods are either fixed or adaptive. Fixed beam formers are fundamentally based on modelled assumptions on the speech signal and the noise field. Based on the assumed model optimal beam formers can be constructed. Optimal function is only guaranteed for perfect model matching. Adaptive beam formers are used to track variations and to compensate for model mismatch. Generally, adaptive beam formers are based on continuous estimates of spatial and statistical information contained in the received speech and noise signal. In general, they are more complex to implement and require complex computations.

**[0006]** It is known to use an array of microphones for speech enhancement by exploiting fundamental properties about spatial and temporal distribution of both the speech and the noise sources. Existing methods for broadband adaptive processing of signals from an array of microphones involve highly complex computing routines, and distortion is introduced in the speech signal. Furthermore they are sensitive to model mismatch. The most common methods also include a voice activity detector (VAD) or a double talk detector (DTD), which will substantially degrade performance due to difficulties in their exact implementation.

**[0007]** Optimal beam formers exist such as the Signal-to-Noise plus Interference Beam former (SNIB) [1], and the Minimum Mean Squares Error Beam former (MMSEB).

**[0008]** For the Signal-to-Noise plus Interference Beam former (SNIB) the output signal-to-noise plus interference ratio (SNIR) is defined as

$$Q = (\text{average signal output power}) / (\text{average noise-plus-interference output power}),$$

and the beam former that maximises the ratio Q, is the optimal Signal-to-Noise plus Interference Beam former (SNIB). The mean signal output power is expressed as a function of the filter weights in the beam former, and the optimal weights, which maximises the output signal-to-noise plus interference ratio Q will have to be defined.

**[0009]** The optimal Minimum Mean Squares Error Beam former (MMSEB) is defined as the beam former that minimizes the mean squares difference between the beam former output when all sources are active, and a single microphone observation, when only the signal of interest is present.

**[0010]** Beam formers can be adequately described both in the time-domain and in the frequency-domain, since the measuring unit of frequency ($s^{-1}$) is the inverse of the measuring unit of time (s).

## Summary of the invention

**[0011]** The invention uses an array of two or more microphones, and the broadband signal from each microphone is passed through a bank of band pass filters separating each broadband signal into several frequency sub-bands. The band pass filtered signals pertaining to the same frequency sub-band are beam formed in adaptive beam formers. Finally, from the adaptively beam formed frequency sub-band signals a single beam formed broadband signal is reconstructed. The sensitivity of the microphone array is thereby focused in a region including the desired sound source, whereby sound signals originating outside that region are suppressed. The beam forming is performed in frequency sub-bands, which requires less computational power than broadband beam forming. The method thus lends itself to the use in mobile communications devices such as mobile telephones.

**[0012]** With the invention the spatial selectivity of the microphone array is beam formed to the speaker, whereby disturbing sound sources are suppressed. Point sources that can be suppressed include speaking persons other than the user, and loudspeakers such as a loudspeaker of a hands free communications device and the loudspeakers of a car stereo system and even moving sources. Diffuse fields that can be suppressed include ambient noise and reverberation in the room.

**[0013]** The invention requires information on the desired signal source. Such information is acquired at times when there is reason to believe that only insignificant noise is present, and the system then records sound signals from the desired sound source alone, eg speech from a speaker, and calculates and stores an estimate of the source correlation matrix and an estimate of the source signal cross correlation vector. The acquisition of the source signal alone can be done in an initial phase where the user selects a mode of operation therefor, or the device using the invention can use eg a voice activity detector to detect voice activity of a speaker, and at instances with only insignificant noise the speaker's voice is recorded and analysed.

## Brief description of the drawing

**[0014]**

Figure 1 shows a schematic block diagram of an apparatus using the method of the invention.

## Detailed description of the invention

**[0015]** In figure 1 a first plurality of *I* microphones M1 - MI are arranged for receiving sound signals from a speaker. The plurality of microphones have fixed positions relative to each other thus forming a fixed array, and their number *I* will be adapted to the actual use. Thus, eg in a handset there will be relatively few microphones such as two, and in a more stationary installation such as in a car a higher number of microphones will typically be used such as four, six or more. In principle any type of microphone can be used. Each microphone outputs an electrical signal representing the sound signal received by the microphone. The electrical output signal from the microphones can be an analogue signal or a digital signal, and in the latter case the microphones will have an analogue-to-digital (A/D) converter.

**[0016]** The output signal of each microphone is input to an individual one of a first plurality of *I* filter banks, which are capable of receiving and processing the signals in analogue or digital form as output by the microphones. In principle, there is one filter bank for each microphone, but by analogue or digital multiplexing methods one filter bank can be used for several microphones.

**[0017]** In principle, all filter banks are identical, and each filter bank has a second plurality of *K* band pass filters where each band pass filter lets a predetermined band of frequencies pass through the filter, so that each filter covers a distinct band of frequencies derived from the microphone output signal. Together the plurality of band pass filters in each bank covers the whole frequency range of interest. Each bank of band pass filters thus outputs a second plurality of *K* band pass filtered signals covering distinct frequency bands.

**[0018]** The construction with *I* filter banks each having *K* band pass filters lends it self to implementation in digital circuits, and in practice the illustrated banks of parallel band pass filters will be implemented using digital signal processing.

**[0019]** A second plurality of *K* beam formers each receive a first plurality of *I* inputs of band pass filtered signals, ie one from each filter bank, where the *I* band pass filtered signals all cover the same frequency band. In each frequency-band the corresponding beam former focuses the sensitivity of the microphone array to a beam or region including the desired sound source, which in this case is a speaking person.

**[0020]** The outputs of the *K* beam formers are fed as input signals to a unit in which a single channel wide frequency-band output signal is reconstructed by combining the *K* beam formed signals.

**[0021]** In the above description the block diagram in figure 1 is used as an illustration for explaining the invention. The invention is preferably implemented in software controlled digital circuits, in which case the illustrated individual

blocks in figure 1 will not be distinct from each other. Rather, the software controlled digital circuits will perform the described operations.

**[0022]** In case of stationary signal conditions and when microphone positions and the speaker's position relative to the microphone array are known and accurate, prior art methods of beam forming may be sufficient, such as the methods described in [3], [4] and [5]. However, when surrounding noise and/or additional disturbing noise sources are changing with time, an adaptive structure or method will perform better and is preferred in order to make use of changing spatial and temporal signal properties.

**[0023]** The invention uses a new algorithm called the Calibrated Weighted Recursive Least Squares (CW-RLS) algorithm. Characteristic features of this algorithm are the introduction of a weighting factor that is used on the observed signal correlation matrix estimates and the use of pre-calculated source correlation estimates. The invention proposes an efficient method of recursively updating estimates.

**[0024]** The MMSE optimal beam former weighting factors in frequency sub-band *k* can be expressed as follows

$$w_{ls,opt}^{(k)}(N) = \arg\min\left(\sum_{n=0}^{N-1}\left[\left|y^{(k)}(n) - s_r^{(k)}(n)\right|^2\right]\right), \quad 1 \le r \le I \tag{1}$$

where *k* is the frequency sub-band number, *N* is the sample number, $y^{(k)}(n)$ is the output of the beam former, and $s_r^{(k)}(n)$ is reference information on the signal source alone. This information is not directly available, or it cannot be expected to be available, and it therefore has to be measured and estimated separately. This is done in a calibration sequence with no or only insignificant background noise, ie the desired signal from the desired signal source such as a speaking person alone. This calibration signal will represent the temporal and spatial information about the desired source. Since the source signal information $S_r^{(k)}(n)$is independent on the actually received sound signals, the Least Squares problem can be separated into two parts:

$$w_{ls,opt}^{(k)}(N) = \arg\min\left(\sum\left[\left|w^{(k)H}s^{(k)}(n) - s_r^{(k)}(n)\right|^2 + \left|w^{(k)H}x^{(k)}(n)\right|^2\right]\right), \quad 1 \le r \le I \tag{2}$$

where the first part can be calculated in advance. Calculating the sum yields

$$w_{ls,opt}^{(k)} = \arg\min\left(w^{(k)H}\left[\hat{R}_{ss}^{(k)}(N) + \hat{R}_{xx}^{(k)}(N)\right]w^{(k)} - w^{(k)H}\hat{r}_s^{(k)}(N) - \hat{r}_s^{(k)H}(N)w^{(k)} + r_{sr}^{(k)}\right) \tag{3}$$

where the estimated source correlation matrix for frequency sub-band *k* can be pre-calculated in the calibration phase as

$$\hat{R}_{ss}^{(k)}(N) = \frac{1}{N}\sum_{n=0}^{N-1} s^{(k)}(n)s^{(k)H}(n), \tag{4}$$

and the estimated source signal spatial cross correlation vector for frequency sub-band *k* as

$$\hat{r}_s^{(k)}(N) = \frac{1}{N}\sum_{n=0}^{N-1} s^{(k)}(n)s_r^{(k)H}(n)$$

where $s^{(k)}(n)=[s_1^{(k)},\ldots,s_I^{(k)}(n)]^T$ are actually received microphone signals when the desired signal source is active alone, ie in the calibration phase. The least squares minimization of equation (3) is found by

$$w_{ls,opt}^{(k)} = \left[\hat{R}_{ss}^{(k)}(N) + \hat{R}_{xx}^{(k)}(N)\right]^{-1} \hat{r}_s^{(k)}(N) \quad (5)$$

where

$$\hat{R}_{xx}^{(k)}(N) = \frac{1}{N}\sum_{n=0}^{N-1} x^{(k)}(n)x^{(k)^H}(n)$$

is the observed/measured signal correlation matrix for frequency sub-band *k*. This means that an estimate of the calibration data can be used in the algorithm.

**[0025]** In the following the Calibrated Weighted Recursive Least Squares (CW-RLS) algorithm is derived. It is a characteristic feature of the CW-RLS algorithm that it introduces a weighting factor on the observed signal correlation matrix estimates, and also uses pre-calculated source correlation estimates in equation (4). The algorithm also achieves this update recursively.

**[0026]** An exponential weighting factor $\lambda$, $0 < \lambda < 1$, which may also be referred to as a "forgetting factor", is introduced in the second part of equation (2) according to

$$w_{ls,opt}^{(k)}(N) = \arg\min\left(\sum_{n=0}^{N-1}\left[\left|w^{(k)^H} s^{(k)}(n) - s_r^{(k)}(n)\right|^2 + \lambda^{N-1-n}\left|w^{(k)^H} x^{(k)}(n)\right|^2\right]\right) \quad (6)$$

where $1 \le r \le l$. An algorithm is thereby obtained that follows the statistical variations in the observed sound signals or data, when the beam former operates in a non-stationary environment. Calculation of the sum gives the same relationship as in equation (3), but the correlation matrix estimates of the observed data will be in accordance with the following equation:

$$\hat{R}_{xx}^{(k)}(N) = \frac{1}{N}\sum_{n=0}^{N-1} \lambda^{N-1-n} x^{(k)}(n)x^{(k)^H}(n) \quad (7)$$

where the least squares solution is given by equation (5). Since the correlation estimates from the calibration sequence, equation (4), are gathered in advance, a recursive update formula for each sample of the new data observation vector $x^{(k)}(n)$ can be formulated.

**[0027]** First, the total correlation matrix is introduced:

$$\hat{R}^{(k)}(n) = \hat{R}^{(k)}_{ss}(N) + \hat{R}^{(k)}_{xx}(n) \quad (8)$$

where it is desired to recursively update the inverse of this matrix. This is done using the Matrix-Inversion Lemma [2]. The total correlation matrix is updated at sample instance *n* according to

$$\begin{aligned}\hat{R}^{(k)}(n) &= \hat{R}_{ss}^{(k)}(N) + \lambda\hat{R}_{xx}^{(k)}(n-1) + x^{(k)}(n)x^{(k)^H}(n) = \\ &\lambda\left[\hat{R}_{ss}^{(k)}(N) + \hat{R}_{xx}^{(k)}(n-1)\right] + x^{(k)}(n)x^{(k)^H}(n) + (1-\lambda)\hat{R}_{ss}^{(k)}(N) = \\ &\lambda\hat{R}^{(k)}(n-1) + x^{(k)}(n)x^{(k)^H}(n) + (1-\lambda)\hat{R}_{ss}^{(k)}(N).\end{aligned} \quad (9)$$

**[0028]** The effect of this updating is that the total correlation matrix is weighted, and that both the rank one "correction term" $x^{(k)}(n)x_H^{(k)}(n)$ and the fraction (1-$\lambda$) of the estimated source correlation matrix, or calibration correlation matrix, $R_{ss}^{(k)}(N)$ multiplied by the weighting factor, are added. This updating can be implemented directly in a two-step procedure

using the Matrix-Inversion Lemma [2], however, this will require complex power and time-consuming calculations of the inverse of the estimated source correlation matrix $R_{ss}^{(k)}(N)$ for each step of updating. Such complex power and time-consuming calculations are undesirable.

**[0029]** One way to circumvent the matrix inversion and thus substantially reduce the complexity of the calculations is to update the total correlation matrix by adding scaled eigenvectors of the estimated source correlation matrix, which will result in several, and simpler, rank one updates as

$$\hat{R}^{(k)}(n) = \lambda \hat{R}^{(k)}(n-1) + x^{(k)}(n)x^{(k)^H}(n) + (1-\lambda)\sum_{p=1}^{I} \gamma_p^{(k)} q_p^{(k)} q_p^{(k)^H} \qquad (10)$$

where $\gamma^{(k)}$ is the p:th eigenvalue, and $q^{(k)}$ is the p:th eigenvector of the I-by-I estimated source correlation matrix $\hat{R}_{ss}^{(k)}(N)$ The weighted optimal recursive least squares solution at sample instant $n$ is then given by

$$w_{ls}^{(k)}(n) = \left[\hat{R}^{(k)}(n)\right]^{-1} \hat{r}_s^{(k)}(N) \qquad (11)$$

where the calibration correlation vector $\hat{r}_s^{(k)}(N)$ is gathered and estimated in advance and is assumed to be uncorrelated with the observed data.

**[0030]** One simple way to further reduce the complexity is sequentially adding one scaled eigenvector at each sample instance. This is easily achieved by replacing the index ρ in equation (10) by the single index ρ = (*n mod I*) + 1. When the statistical properties of the environmental noise change abruptly, eg when a new source of disturbance suddenly appears, a smoothing of the weights may be appropriate. A first order auto regressive (AR) model is preferred for the smoothing, and the weight update then becomes

$$w_{ls}^{(k)}(n) = \alpha w_{ls}^{(k)}(n-1) + (1-\alpha)\left[\hat{R}^{(k)}(n)\right]^{-1} \hat{r}_s^{(k)}(N) \qquad (12)$$

where α is the AR-parameter, which corresponds to the real-valued pole of the AR-model. By applying the Matrix-Inversion Lemma in order to update the inverse of the total correlation matrix an efficient implementation of the algorithm of the invention is obtained.

**[0031]** The method of beam forming according to the invention has two main phases. In the first phase, information on the desired signal source alone is gathered. The first phase will be referred to as the calibration phase. The second phase is referred to as the operation phase.

**[0032]** In the calibration phase, in principle, the desired signal source such as a speaking person is active alone. With the arrangement in figure 1 sound signals from a speaking person are captured by the array of microphones. Output signals from the first plurality *I* microphones are filtered in the filter banks, whereby for each microphone a second plurality *K* of band pass filtered signals covering distinct frequency bands are derived from the microphone signals. For each of the distinct frequency sub-bands the estimated source correlation matrix $R_{ss}^{(k)}$ is calculated and stored in a memory. This is done simultaneously for all frequency sub-bands. When there are other known disturbing or undesired sound sources, eg a loudspeaker for hands-free operation of the telephone, with a fixed position relative to the microphone array, correlation estimates from these signals are added to the stored estimated source correlation matrix $R_{ss}^{(k)}$ An estimated source signal spatial cross correlation vector $\hat{r}_s^{(k)}$ for frequency sub-band *k* is also calculated and stored.

**[0033]** The calibration may be performed in an initial phase and in quiet environments, but the system will preferably perform the calibration also during the operation phase at times where there are reasons to believe that undesired sounds such as background noise are of minor importance and have a negligible influence on the estimation of the source correlation matrix. Thereby the estimated source correlation matrix $R_{ss}^{(k)}$ is updated currently. Methods of detecting such times for updating the estimated source correlation matrix are known and involve eg a voice activity detector (VAD).

**[0034]** The calibration is performed as follows. With the desired signal source, eg a speaking person, as the only or at least the dominating sound source, the microphone signal vector

$$x^{(k)}(n) = [x_1^{(k)}(n), \ldots, x_I^{(k)}(n)]^T$$

is calculated for each of the $K$ distinct frequency sub-bands. The estimated source signal spatial cross correlation vector for each frequency sub-band $k$ and for each sample $n$ is calculated as follows:

$$\hat{r}_s^{(k)}(N) = \frac{1}{N}\sum_{n=1}^{N} x^{(k)}(n)x_r^{(k)^H}(n),$$

and the estimated source correlation matrix is calculated as follows:

$$\hat{R}_{ss}^{(k)}(N) = \frac{1}{N}\sum_{n=1}^{N} x^{(k)}(n)x^{(k)^H}(n).$$

With all known undesired disturbing sound sources being active the corresponding microphone signal vector

$$x^{(k)}(n)=[x_1^{(k)}(n)....,x_I^{(k)}(n)]^T$$

is calculated for each of the $K$ distinct frequency sub-bands, and correspondingly the estimated undesired disturbance correlation matrix is calculated as follows:

$$\hat{R}_{dd}^{(k)}(N) = \frac{1}{N}\sum_{n=1}^{N} x^{(k)}(n)x^{(k)^H}(n).$$

**[0035]** The correlation matrices are stored in diagonalized form:

$$\left(\hat{R}_{dd}^{(k)}(N) + \hat{R}_{ss}^{(k)}(N)\right) = Q^{(k)^H}\Gamma Q^{(k)}$$

**[0036]** The eigenvectors are denoted

$$Q^{(k)}= [q_1^{(k)},...,q_I^{(k)}],$$

and the eigenvalues are denoted

$$\Gamma^{(k)} = diag\left(\left[\gamma_1^{(k)},...,\gamma_I^{(k)}\right]\right).$$

**[0037]** For each frequency sub-band $k$, the eigenvectors $q_i^{(k)}$, the eigenvalues $\gamma_i^{(k)}$, and the cross correlation vector $\hat{r}_s^{(k)}$ (N), where $0 \leq k \leq K-1$, are stored in memory for subsequent use in the operation phase.

**[0038]** In the operation phase it is thus assumed that, for each of the $K$ distinct frequency sub-bands, the estimated source correlation matrix $\hat{R}_{ss}^{(k)}$ and the estimated source signal spatial cross correlation vector $\hat{r}_s^{(k)}$ are stored and available, either from a separate initial calibration phase or from a more recent updating.

**[0039]** In each of the $K$ frequency sub-bands, the following variables are used:

- beam former weighting vectors:

$$w_n^{(k)} =[w_1^{(k)}(n),...,w_I^{(k)}(n)]^T$$

- the inverse of the total correlation matrix variable at time instant $n$, for frequency sub-band number $k$: $P_n^{(k)}$, Initialize

as:

$$P_0^{(k)} = Q^{(k)^H} \Gamma^{-1} Q^{(k)}$$

- the forgetting factor $\lambda$ for the WRLS algorithm and a weight smoothing factor $\alpha$ for the weighting factor update, which are preferably both chosen as constants for all frequency sub-bands:

$$\lambda^{(k)}=\lambda,\ \alpha^{(k)} = \alpha.$$

**[0040]** The algorithm is as follows. When any two or more of the sound sources (speech and noise) are active simultaneously, the following quantities are computed for each sample *n* and for each frequency sub-band *k*:

$$x_n^{(k)} = [x_1^{(k)}(n),\ldots,x_l^{(k)}(n)]^T$$

$$P_n^{(k)} = \lambda^{-1} P_{n-1}^{(k)} - \frac{\lambda^{-2} P_{n-1}^{(k)} x_n^{(k)} x_n^{(k)^H} P_{n-1}^{(k)}}{1+\lambda^{-1} x_n^{(k)^H} P_{n-1}^{(k)} x_n^{(k)}}$$

$$P_n^{(k)} = P_n^{(k)} - \frac{\gamma_p(1-\lambda)\, P_n^{(k)}\, q_p^{(k)} q_p^{(k)^H} P_n^{(k)}}{1+\gamma_p(1-\lambda) q_p^{(k)^H} P_n^{(k)}\, q_p^{(k)}}$$

where index $p=(n \bmod l)+1$,

$$w_n^{(k)} = \alpha w_{n-1}^{(k)} + (1-\alpha) P_n^{(k)} \hat{r}_s^{(k)}.$$

**[0041]** For each frequency sub-band *k* the output from the corresponding beam former then is:

$$y^{(k)}(n) = w_n^{(k)^H} x_n^{(k)}.$$

**[0042]** The algorithm is then repeated for the next sample n+1, etc.

**[0043]** In the operation phase the microphone output signals are continuously decomposed into discrete frequency sub-bands. The sub-band weighting factors $w^{(k)}$ are updated by making use of both the stored correlation estimates and of the actual microphone observations. The totality of beam former output signals, which each is a beam formed frequency sub-band time-domain signal, are input to a reconstruction filter bank. The output of the reconstruction filter bank is taken as the estimate of the sound signal from the desired sound source. Once the correlation estimates are stored in the memory, the algorithm is continuously adapting.

**[0044]** The algorithm contains a step in which a rank one update of the correlation matrix is performed using scaled eigenvectors, one eigenvector for each new input data vector. This step adds correlation estimates from the source signal, whereby information gathered in the acquisition or calibration phase will remain a constant part of the correlation matrix, while the contributions form the undesired environmental noise will be subject to the forgetting factor $\lambda$ in the estimates.

**[0045]** Good performance of the algorithm of the invention requires that the number *K* of frequency sub-bands is large enough for the frequency-domain representation to be accurate. In other words, the number of frequency sub-bands is proportional to the length of the equivalent time-domain filters (filter length = the number of parameters used in a digital filter), and the number of degrees of freedom in the beam formers increases with the number of frequency sub-bands. Also, the delay caused by the frequency transformations is related to the number of frequency sub-bands.

**[0046]** Time-domain and frequency-domain representations are closely related, and the algorithm of the invention can easily be extended to a combination of time-domain and frequency-domain representations. Each frequency sub-band signal can also be regarded as a time-domain signal samples at a reduced sampling rate, ie proportional to the frequency sub-band bandwidth, and having substantially only the frequencies in the sub-band. By applying the time-

domain algorithm in each frequency sub-band, the degrees of freedom for the band-pass filters are increased, while the number of sub-bands may be kept constant. The lengths of the sub-band filters may differ between sub-bands, and the consequence is that a multi-resolution sub-band identification is obtained.

**[0047]** The extension of the algorithm is achieved simply by using more lags from the observed microphone signals $x_i^{(k)}(n)$, $1 \le i \le I$, when defining the input vector

$$x^{(k)} = [x_1^{(k)}, \ldots, x_I^{(k)}]^T$$

where each element consists of $L^{(k)}$ lags as

$$x_i^{(k)} = [x_i^{(k)}(n), x_i^{(k)}(n-1), \ldots, x_i^{(k)}(n-L^{(k)}+1)],\ 1 \le i \le I$$

and the weight factors for each sub-band are similarly extended as

$$w^{(k)} = [w_1^{(k)}, \ldots, w_I^{(k)}]^T$$

where each element consists of $L(k)$ parameters

$$w_i^{(k)} = [w_i^{(k)}(0), w_i^{(k)}(1), \ldots, w_i^{(k)}(L^{(k)}-1)],\ 1 \le i \le I.$$

**[0048]** The definitions of the correlation matrix, the eigenvalue and eigenvector matrices follow directly, and the size of the matrices will be increased by the factor $L^{(k)} \cdot L^{(k)}$ for frequency sub-band number $k$. The amount of memory needed to store the eigenvectors and the eigenvalues increases with increasing number of sub-band used.

**[0049]** The band-pass filtering or decomposition of the broadband microphone signals into frequency sub-bands is preferably done using a uniform Discrete Fourier Transform (DFT), eg as described in [6], to decompose the full-rate sampled signals $x_i(n)$ into $K$ sub-band signals. The sub-bands are preferably created in such a way that a prototype filter with a low-pass characteristic is used to ensure that the response from the $k$-th sub-band is the same as that of the prototype filter, although centred at a normalized frequency $2\pi k/K$, whereby the set of $K$ sub-bands will cover the whole frequency range. In a modulated filter bank the prototype filter equals one of the filters in the bank (usually the first filter), and the other filters are modulated versions of the prototype filter. Thus only the prototype filter needs to be created. Each sub-band signal is thereby represented in the base band. The filter bank should cover the entire frequency range, and a redundant, ie over-complete, frequency sub-band decomposition and reconstruction should be used. The invention is not limited to using uniformly distributed frequency sub-bands or a modulated filter bank.

**[0050]** In order to reduce aliasing between sub-bands, the sub-band decomposition is made over-sampled.

**[0051]** In the reconstruction filter bank a time domain signal is reconstructed or synthesized from the beam formed frequency sub-band signals from the beam formers. The beam formed frequency sub-band signals are up-converted from the base band to the actual frequency band, and summed in the reconstruction filter.

**[0052]** In practical use, eg for hands-free operation of a mobile telephone in a car, the microphone array can eg have six microphones in a linear configuration with a spacing of 50 mm between microphones, and the microphone array can be placed at a nominal distance of eg 350 mm from the speaker's mouth. The number $K$ of band pass filters in each filter bank can be in the range from 32 or less to 256 or more, typically 64.

**List of terms used**

**[0053]**

| | |
|---|---|
| $K$ | = total number of frequency sub-bands |
| $k$ | = index number of frequency sub-band, $0 \le k \le K-1$ |
| $I$ | = number of microphone channels, ie number of microphones in the array |
| $i$ | = microphone index in the array, $1 \le i \le I$ |
| $N$ | = number of samples |
| $n$ | = time signal sample number, sample instant |
| $x_i^{(k)}(n)$ | = observed signal from microphone $i$ in frequency sub-band $k$ at sample $n$ |
| $y$ | = output signal from beam formers |

$\vec{a}$ $e^{j2\pi f}$

$\hat{R}^{(k)}$ = estimated source correlation matrix for frequency sub-band *k*

$\hat{r}^{(xs)}$ = estimated source signal spatial cross correlation vector for frequency sub-band *k*

$\hat{R}^{(k)}$ = observed/measured signal correlation matrix for frequency subband *k*

$R^{(nn)}$ = estimated noise (disturbance) correlation matrix for frequency subband *k*

$w^{(k)}$ = beam former weighting factors for frequency sub-band *k*

$\lambda$ $^{ls}$ = forgetting factor, $0 < \lambda < 1$, typical value: $\lambda = 0.99$

$\alpha$ = weight smoothing factor, typical value: $\alpha = 0.01$

$q_i$ = eigenvector

$\gamma_p^{(k)}$ = *p*:th eigenvalue

**References**

**[0054]**

[1] J. E. Hudson, "Adaptive Array Principles", Peter Peregrinus Ltd., 1991, ISBN 0-86341-247-5.

[2] S. Haykin, "Adaptive Filter Theory", Prentice Hall International, Inc., 1996, ISBN 0-13-397985-7.

[3] M. M. Goulding, J. S. Bird, "Speech Enhancement for Mobile Telephony", IEEE Transactions on Vehicular Technology, vol. 39, no. 4, pp. 316-326, November 1990.

[4] S. Nordholm, V. Rehbock, K. L. Toe, S. Nordebo, "Chebyshev Optimization for the Design of Broadband Beam formers in the Nearfield", IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing, vol. 45, pp. 141-143, 1998.

[5] S. Nordholm, Y. H. Leung, "Performance Limits of the Generalized Sidelobe Cancelling Structure in an Isotropic Noise Field", Journal of the Acoustical Society of America, vol. 107, no. 2, pp. 1057-1060, February 2000.

[6] Mitra, Sanjit Kumar, "Digital Signal Processing", McGraw-Hill Companies Inc., 1998, ISBN 0-07-042953-7.

**Claims**

1. A method of enhancing received desired sound signals from a desired sound source and of suppressing received undesired sound signals from one or more undesired sound sources, the method comprising

   - receiving, at different distinct locations, a first plurality (I) of sound signals and converting each of the received sound signals into a corresponding electrical signal representing an individual one of the received sound signals,
   - deriving, from each of the electrical signals representing individual ones of the received sound signals, a second plurality (K) of signals representing band pass filtered signals each covering a distinct frequency band
   - for each of the distinct frequency bands, beam forming the second plurality (K) of band pass filtered signals covering the corresponding distinct frequency band to have a maximum sensitivity at a region including the desired sound source while attenuating the undesired signals, and
   - combining the beam formed band pass filtered signals to an output signal.

2. A method according to claim 1, wherein the beam forming is based on temporal and/or spatial information on the desired sound source.

3. A method according to claim 1, wherein the beam forming is based on temporal and/or spatial information on the undesired sound source or sources.

4. A method according to claim 2, wherein, for each of the second plurality (K) of distinct frequency bands, an estimated source correlation matrix

$$\hat{R}_{xx}^{(k)}(N) = \frac{1}{N}\sum_{n=1}^{N} x^{(k)}(n)x^{(k)^H}(n)$$

and an estimated source signal cross correlation vector

$$\hat{r}_s^{(k)}(N) = \frac{1}{N}\sum_{n=1}^{N} x^{(k)}(n)x_r^{(k)^H}(n)$$

are calculated and stored,
where
$x^{(k)}$ is the signal received at a reference position and band pass filtered in frequency band number $k$, and
$x^{(k)}(n) = [x_1^{(k)}(n),...,x_l^{(k)}(n)]^T$ is a received reference signal vector in frequency band number $k$, $0 \le k \le K-1$, at sample instant $n$.

5. A method according to any one of claims 3-4, wherein, for each of the second plurality (K) of distinct frequency bands, an estimated interference correlation matrix

$$\hat{R}_{dd}^{(k)}(N) = \frac{1}{N}\sum_{n=1}^{N} x^{(k)}(n)x^{(k)^H}(n)$$

is calculated and stored.

6. A method according to any one of claims 1-5, wherein beam former matrix weighting factors are updated exponentially in time.

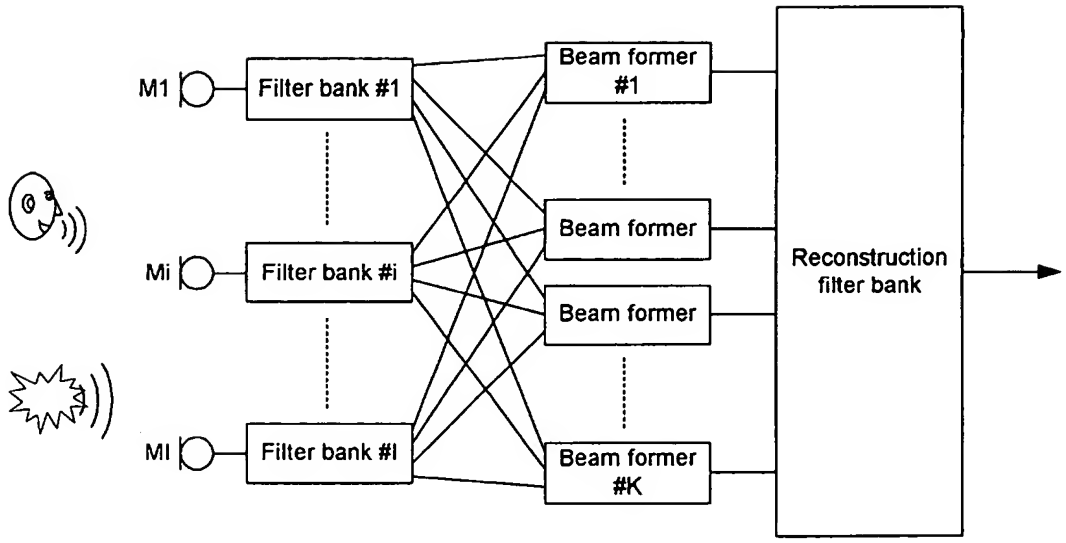
M1
Filter bank #1
Mi
Filter bank #i
Ml
Filter bank #l
Beam former #1
Beam former
Beam former
Beam former #K
Reconstruction filter bank

Fig. 1

European Patent Office

# EUROPEAN SEARCH REPORT

Application Number
EP 02 38 8021

## DOCUMENTS CONSIDERED TO BE RELEVANT

| Category | Citation of document with indication, where appropriate, of relevant passages | Relevant to claim | CLASSIFICATION OF THE APPLICATION (Int.Cl.7) |
|---|---|---|---|
| X | PATENT ABSTRACTS OF JAPAN<br>vol. 2002, no. 06,<br>4 June 2002 (2002-06-04)<br>& JP 2002 062348 A (SONY CORP),<br>28 February 2002 (2002-02-28)<br>& US 2002/048376 A1 (UKITA MASAKAZU)<br>25 April 2002 (2002-04-25)<br>* page 10, left-hand column, line 45 - page 11, left-hand column, line 58; figures 9,10 *<br>--- | 1-3 | H04R3/00<br>H04R1/40<br>G10K11/178 |
| X | SYDOW C: "BROADBAND BEAMFORMING FOR A MICROPHONE ARRAY"<br>JOURNAL OF THE ACOUSTICAL SOCIETY OF AMERICA, AMERICAN INSTITUTE OF PHYSICS. NEW YORK, US,<br>vol. 96, no. 2, PART 1,<br>1 August 1994 (1994-08-01), pages 845-849, XP000466113<br>ISSN: 0001-4966<br>* page 845, right-hand column, line 35 - page 846, left-hand column, line 9 *<br>--- | 1-3 | TECHNICAL FIELDS SEARCHED (Int.Cl.7)<br>H04R<br>G10K |
| A | DAHL M ET AL: "ACOUSTIC NOISE AND ECHO CANCELING WITH MICROPHONE ARRAY"<br>IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, IEEE INC. NEW YORK, US,<br>vol. 48, no. 5, September 1999 (1999-09), pages 1518-1526, XP000912523<br>ISSN: 0018-9545<br>* the whole document *<br>---<br>-/-- | 1-6 | |

The present search report has been drawn up for all claims

| Place of search | Date of completion of the search | Examiner |
|---|---|---|
| MUNICH | 7 November 2002 | Lindberg, P |

CATEGORY OF CITED DOCUMENTS

X : particularly relevant if taken alone
Y : particularly relevant if combined with another document of the same category
A : technological background
O : non-written disclosure
P : intermediate document

T : theory or principle underlying the invention
E : earlier patent document, but published on, or after the filing date
D : document cited in the application
L : document cited for other reasons

& : member of the same patent family, corresponding document

EPO FORM 1503 03.82 (P04C01)

European Patent Office

# EUROPEAN SEARCH REPORT

Application Number
EP 02 38 8021

## DOCUMENTS CONSIDERED TO BE RELEVANT

| Category | Citation of document with indication, where appropriate, of relevant passages | Relevant to claim | CLASSIFICATION OF THE APPLICATION (Int.Cl.7) |
|---|---|---|---|
| A | LLEIDA E ET AL: "Robust continuous speech recognition system based on a microphone array" ACOUSTICS, SPEECH AND SIGNAL PROCESSING, 1998. PROCEEDINGS OF THE 1998 IEEE INTERNATIONAL CONFERENCE ON SEATTLE, WA, USA 12-15 MAY 1998, NEW YORK, NY, USA,IEEE, US, 12 May 1998 (1998-05-12), pages 241-244, XP010279154 ISBN: 0-7803-4428-6 * page 242, right-hand column, line 22 - line 38 * ----- | 1-6 | |

TECHNICAL FIELDS SEARCHED (Int.Cl.7)

The present search report has been drawn up for all claims

| Place of search | Date of completion of the search | Examiner |
|---|---|---|
| MUNICH | 7 November 2002 | Lindberg, P |

CATEGORY OF CITED DOCUMENTS

X : particularly relevant if taken alone
Y : particularly relevant if combined with another document of the same category
A : technological background
O : non-written disclosure
P : intermediate document

T : theory or principle underlying the invention
E : earlier patent document, but published on, or after the filing date
D : document cited in the application
L : document cited for other reasons

& : member of the same patent family, corresponding document

EPO FORM 1503 03.82 (P04C01)

## ANNEX TO THE EUROPEAN SEARCH REPORT ON EUROPEAN PATENT APPLICATION NO. EP 02 38 8021

This annex lists the patent family members relating to the patent documents cited in the above-mentioned European search report. The members are as contained in the European Patent Office EDP file on
The European Patent Office is in no way liable for these particulars which are merely given for the purpose of information.

07-11-2002

| Patent document cited in search report | Publication date | Patent family member(s) | Publication date |
|---|---|---|---|
| JP 2002062348 A | 28-02-2002 | US 2002048376 A1 | 25-04-2002 |

EPO FORM P0459

For more details about this annex : see Official Journal of the European Patent Office, No. 12/82